Available online at www.sciencedirect.com

ScienceDirect

Procedia Engineering 145 (2016) 518 – 524

Procedia Engineering

www.elsevier.com/locate/procedia

International Conference on Sustainable Design, Engineering and Construction

Text analytics for supporting stakeholder opinion mining for large-scale highway projects

Xuan Lv[a], Nora El-Gohary[b]*

[a]*Graduate Student, Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, 205 North Mathews Ave., Urbana, IL 61801, United States*

[b]*Assistant Professor, Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, 205 North Mathews Ave., Urbana, IL 61801, United States*

Abstract

For large-scale highway projects, late identification of stakeholder concerns often leads to design changes and duplication of effort, which may cause major project delays. This paper proposes a stakeholder opinion mining approach for helping transportation practitioners better identify the types of concerns in the early project stage. The proposed approach includes two major components: (1) stakeholder concern extraction, and (2) stakeholder concern classification. This paper focuses on presenting the proposed methodology and experimental results for stakeholder concern extraction, which extracts the words and phrases that describe stakeholder concerns from stakeholder comments on large-scale highway projects. In developing the proposed stakeholder concern extraction methodology, several supervised machine learning (ML) algorithms were tested and evaluated, and the effect of using a predefined name list as feature was also investigated. All the algorithms were tested on a testing data set of 200 comment sentences, which were selected from a comment collection including 1,849 stakeholder comments on five large-scale highway projects.

Peer-review under responsibility of the organizing committee of ICSDEC 2016

Keywords: Data analytics; Opinion mining; Information extraction; Natural language processing; Machine learning.

* Corresponding author. Tel.: +1-217-333-6620; fax: +1-217-265-8039.
E-mail address: gohary@illinois.edu

Peer-review under responsibility of the organizing committee of ICSDEC 2016
doi:10.1016/j.proeng.2016.04.039

1. Introduction

Because large-scale highway projects tend to have significant impact on the surrounding natural environment, the everyday life of the affected public, and the regional economic development, they are required to undergo a comprehensive environmental review, during which the opinions of a wide spectrum of different stakeholders ranging from government agencies to the general public are solicited. Transportation agencies spend a large amount of time and money on the environmental review process. For example, the median time to complete the environmental review process for large-scale highway projects was over 7 years in 2013, and the cost can be several million dollars [1]. Despite that, late identification of stakeholder concerns has been identified as one of the major causes for the lengthy and costly project development process [2]. There is, thus, a need for identifying stakeholder concerns in the early project stage to improve the efficiency of transportation decision making. To address this need, this paper proposes a stakeholder opinion mining approach, which consists of two major components: stakeholder concern extraction and stakeholder concern classification. Stakeholder concern extraction aims at extracting words and phrases that describe stakeholder concerns from stakeholder comments on large-scale highway projects. Stakeholder opinion classification aims at classifying the extracted concern words and phrases into different concern types. This paper focuses on describing the proposed methodology and experimental results for stakeholder concern extraction.

2. Background and knowledge gaps

A stakeholder concern is an issue that is affected, positively or negatively, by the project, such as property value, farmland, fuel tax, population growth, and nearby environmental resources. The concern extraction problem can be defined as an aspect extraction problem. Aspect extraction problems have long been studied in the field of aspect-based opinion mining. There are three main aspect extraction approaches that have been proposed in recent years: (1) language rule-based approach, (2) supervised machine learning (ML)-based approach, and (3) topic model-based approach.

The language rule-based approach extracts aspects using predefined rules, which capture the contextual patterns and/or grammatical relations between the terms in the text [3]. Hu and Liu [4] proposed an extraction method based on association rules, which (1) find frequent aspects through frequent nouns and noun phrases, and (2) identify infrequent aspects using relations between aspects and opinion words. The relationships of opinion words and aspects can be generalized as dependency relations, which are widely used rules for aspect extraction. For example, Qiu et al. [5] developed the double propagation methods to extract aspects and opinions simultaneously based on direct dependency relations. Poria et al. [6] exploited common-sense knowledge and sentence-dependency trees to detect both explicit and implicit aspects from product reviews. One limitation of the rule-based approach is the adaptability of language rules, because the performance of rules depends largely on the type of collection; rules that work well on one collection may not work well on another collection.

The supervised ML-based approach learns to extract aspects from manually labeled data. Some methods utilized sequence models, which treat aspect extraction as a sequence-labeling task. For example, Jin et al. [7] utilized lexicalized hidden Markov model (HMM), which incorporated linguistic features such as part-of-speech and lexical patterns to extract aspects from product reviews. Jakob and Gurevych [8] evaluated the performance of a conditional random field (CRF)-based method for aspect extraction in a single and cross-domain environment. Shariaty and Moghaddam [9] also employed CRF for identifying product aspects and proposed a technique for defining and filtering features to enhance the performance. Other researchers used other supervised learning models that treat aspect extraction as a binary or multi-class classification task. For example, Ghani et al. [10] used both supervised and semi-supervised algorithms to extract attribute and value pairs from product descriptions. Yu et al. [11] trained a one-class support vector machine (SVM) to identify aspects in the candidate noun phrases extracted from Pros and Cons consumer reviews.

The topic model-based approach assumes that the comments are generated though mixtures of topic models, and each topic model is a unigram language model that represents a type of aspect. Mukherjee and Liu [12] developed two joint aspect-opinion models for extracting and categorizing aspects at the same time given user-provided seed words. Chen et al. [13] proposed an aspect extraction framework to extract more coherent aspects by exploiting the

knowledge automatically learned from online reviews. One major limitation of the topic model-based approach is that it can only find some general aspects, and has difficulty in finding fine-grained or precise aspects.

Although a number of aspect-based opinion mining studies have been conducted, two limitations have been identified: (1) most of the research efforts focus on stakeholder opinions on products or services, which are different from stakeholder opinions on large-scale highway projects in terms of the concerns expressed; (2) most of the research efforts focus on one stakeholder group (e.g., users of the product/service), while there are multiple stakeholder groups (e.g., resource agencies, residents, land owners) in the stakeholder comments for large-scale highway projects.

To address the above-mentioned limitations, this paper proposes a stakeholder concern identification approach, which extracts concern words and phrases from stakeholder comments on large-scale highway projects, and classifies the extracted concern words and phrases into different concern types. In preparing the comment collection, the comments of five stakeholder groups were collected, including federal agencies, state agencies, local governments, public organizations, and interested individuals. In developing the concern extraction methodology, the performance of four ML algorithms were evaluated: HMM, CRF, maximum entropy (ME), and multi-class SVM algorithms. To further improve the concern extraction performance, the effect of using a predefined name list was also evaluated.

3. Proposed stakeholder concern extraction methodology

3.1. Data collection

To create a comment collection, five large-scale highway projects from five states were identified (Table 1). For these projects, the comments for all stakeholder groups (federal agencies, state agencies, local governments, public organizations, and interested individuals) that were received during the public comment period, including comments submitted through the project websites, emails, and public hearings, were gathered into a comment collection. The comment collection contains 1,849 comments.

Table 1. Identified large-scale highway projects

Project name	Project location	Number of comments
Cleveland opportunity corridor	Ohio	136
I-395 transportation system	Maine	134
US281	Texas	641
I-5 north coast corridor	California	339
North I-25	Colorado	599
Total	NA	1,849

3.2. Data preparation

To facilitate the implementation of ML algorithms, the Begin-Inside-Outside (BIO) labeling schema was adopted when extracting concerns manually and automatically. In the BIO schema, B stands for beginning of a concern, I stands for inside of a concern, and O stands for outside of a concern. For example, a sample comment and its standard labels are shown in Fig.1.

A total of 200 comments out of the collection were randomly selected as the training data, which include 1,012 sentences. Another 50 comments were randomly selected as the testing data, which include 200 sentences. To create the gold standard, the comments in the training and testing data were manually examined and annotated. The standard labels of each comment were determined based on mutual agreement of the three annotators (the first author and another two researchers).

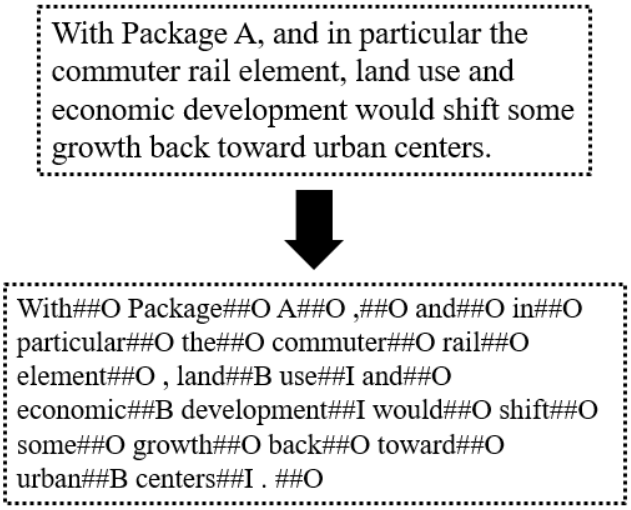

Fig. 1. A sample comment and its standard labels.

3.3. Machine learning algorithm selection

A set of supervised ML algorithms that are commonly used for information extraction were selected for implementation, including HMM, CRF, ME, and SVM.

HMM is a probabilistic generative model for sequential data [3]. In the context of concern extraction, each comment sentence is regarded as a sequence of features, and the target output is a hidden sequence of labels. An HMM model assumes that the current label only depends on its previous label, and that the current features only depend on the current label.

One limitation of HMM is that its assumptions may not be valid in the scenario of concern extraction task. Therefore, another sequence model CRF, which directly models the conditional probability of the target labels given the observed features, was used.

The concern extraction problem can also be formulated as a multi-class classification task. Therefore, the ME and the multi-class SVM algorithms were also implemented. The ME algorithm is based on the ME principle that among all the distributions that satisfy feature constraints, the one with the highest entropy should be selected [14]. One advantage of the ME algorithm is that it does not make any assumptions of conditional independency among features.

The multi-class SVM algorithm implemented in the paper adopted the most commonly-used one-vs-all strategy, which trains a classifier for each class, and assigns the class with the highest confidence score.

The following types of features were used as the input for the above mentioned ML algorithms:

- Token: This feature represents the current term;
- Part of speech (POS): This feature represents the POS tag of the current term;
- Lemma: This feature represents the lemma of the current term;
- Relation head: This feature represents whether the current term is the head of the selected dependency relation. Three different dependency relations were considered and their information were obtained from the Stanford dependency parser [15];
- Relation dependent: This feature represents whether the current term is the dependent of the selected dependency relation.
- Head: This feature represents the head of the current term in the dependency tree;
- Stop-word: This feature represents whether the current term is a strop-word.

3.4. Evaluation

The ML algorithms were evaluated using precision, recall, and F1 measure. Precision, here, is defined as the ratio of the number of correctly extracted concern entities over the total number of extracted entities. Recall, here, is defined as the ratio of the number of correctly extracted concern entities over the total number of concern entities that should be extracted. F1 is the harmonic mean of precision and recall. These measures were calculated based on a comparison of the experiment results with the manually-developed gold standard, for the extracted concern entities from the set of randomly-selected comments in the comment collection.

3.5. Performance improvement

A stakeholder concern name list that contains terms of potential stakeholder concerns was used to further improve the performance of concern extraction. Ratinov and Roth [16] indicated that name lists were useful in similar information extraction tasks (name entity recognition). In order to develop the stakeholder concern name list, a stakeholder concern hierarchy was developed based on a domain semantic model [17]. A partial view of the "stakeholder concern" hierarchy is shown in Fig. 2. The stakeholder concern name list was created including every concept in the stakeholder concern hierarchy, and a new feature was created to represent whether the current term appears in the stakeholder concern name list or not.

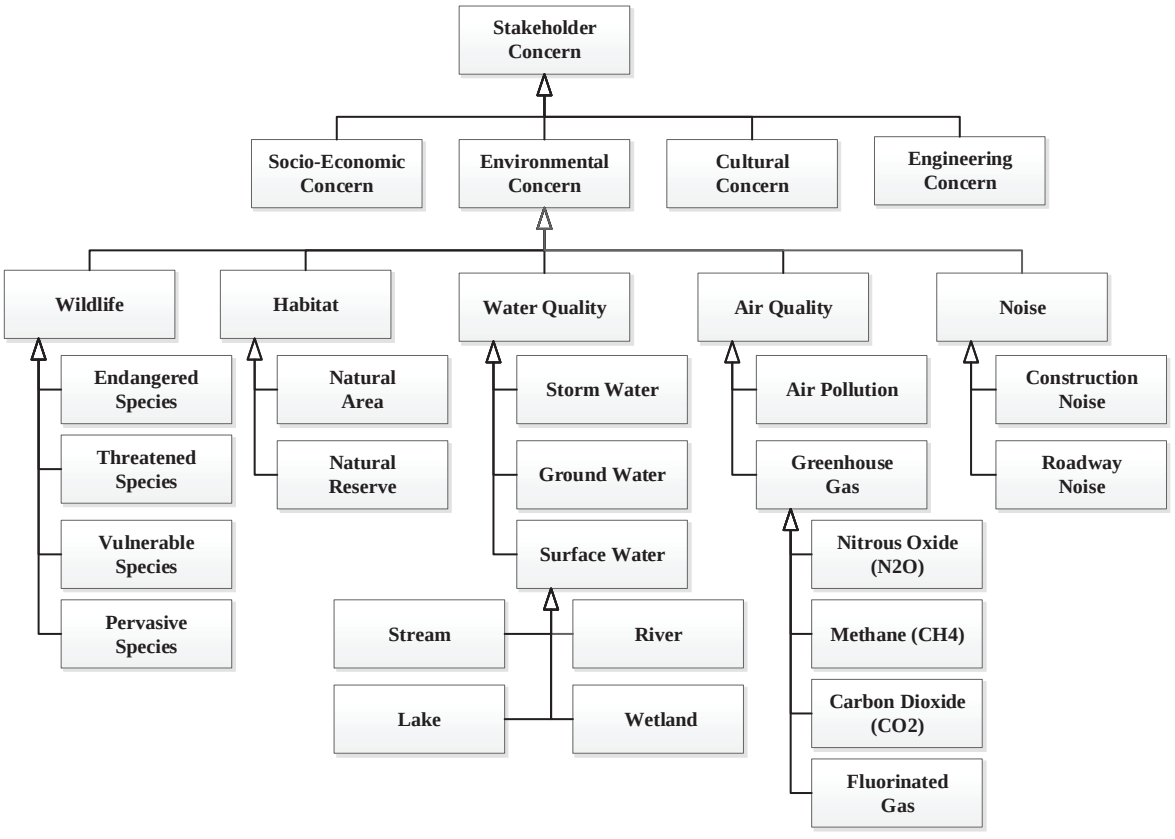

Fig. 2. a partial view of the "stakeholder concern" hierarchy.

4. Experimental results and analysis

The performance of the four algorithms are summarized in Table 2. The CRF algorithm achieved the best performance with 79% precision, 72% recall, and 75% F1 measure. The CRF algorithm outperformed the other algorithms for the following reasons: (1) compared with HMM, it provided some relaxation to the independency assumptions that may not be valid; (2) it considered label-to-label and feature-to-label dependency relations which are important in concern extraction, and were overlooked by the ME and the multi-class SVM algorithms. Therefore, the CRF algorithm was selected to further implement and test the performance of improvement strategy.

Table 2. Performance of the four algorithms

Algorithm	Precision	Recall	F1 measure
HMM	71%	67%	69%
CRF	79%	72%	75%
ME	72%	66%	69%
Multi-class SVM	75%	70%	72%

As shown in Table 3, after using the stakeholder concern name list as a feature, the precision was improved from 79% to 82%, the recall was improved from 72% to 76%, and the F1 measure was improved from 75% to 79%.

Table 3. Performance before and after using the stakeholder concern name list

Features	Precision	Recall	F1 measure
Original features	79%	72%	75%
Original features + name list feature	82%	76%	79%

5. Conclusion and future work

This paper presented a stakeholder opinion mining approach for identifying stakeholder concerns early in the project stage. The proposed approach includes concern extraction and concern classification. In developing the concern extraction method, a set of supervised ML algorithms were tested, and the effect of using a stakeholder concern name list was investigated. All the algorithms were tested on a testing data set of 200 comment sentences and evaluated in terms of precision, recall, and F1 measure. Based on the evaluation, the final algorithm achieved 82% precision, 76% recall, and 79% F1 measure.

In the future, the author will continue to improve the current work in three directions: (1) evaluate more supervised ML algorithms, such as the structural SVM algorithm, to improve the performance of concern extraction; (2) implement feature selection techniques to further improve the performance of concern extraction; and (3) explore or develop algorithms for classifying the extracted concerns into different stakeholder concern groups.

Acknowledgements

This paper is based upon work supported by the Strategic Research Initiatives (SRI) Program by the College of Engineering at the University of Illinois at Urbana-Champaign.

References

[1] U.S. Government Accounting Office (USGAO). *Many Federal and State Review Requirements are Similar, and Little Duplication of Efforts Occurs*. Washington, D.C.: USGAO; 2014, p. 9-12.

[2] Mallet WJ, Luther L. Accelerating Highway and Transit Project Delivery: Issues and Options for Congress. Washington, D.C.: Congressional Research Service; 2011, p. 10-18.

[3] Zhang L, Liu B. Aspect and entity extraction for opinion mining. *Data mining and knowledge discovery for big data*. Berlin, Heidelberg: Springer; 2014, p. 1-40
[4] Hu M, Liu B. Mining opinion features in customer reviews. *Proceedings of National Conference on Artificial Intelligence* 2004; 755-760.
[5] Qiu G, Liu B, Bu J, Chen C. Opinion word expansion and target extraction through double propagation. *Computational linguistics* 2011; 37(1): 9-27.
[6] Poria, S, Cambria, E, Ku, LW, Gui, C, Gelbukh, A. A rule-based approach to aspect extraction from product reviews. *Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP)* 2014; 28-37.
[7] Jin W, Ho HH, and Srihari RK. A novel lexicalized HMM-based learning framework for web opinion mining. *Proceedings of the 26th Annual International Conference on Machine Learning* 2009; 465-472.
[8] Jakob N, Gurevych I. Extracting opinion targets in a single-and cross-domain setting with conditional random fields. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* 2010; 1035-1045.
[9] Shariaty, S, Moghaddam S. Fine-grained opinion mining using conditional random fields. *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on* 2011; 109-114.
[10] Ghani, R, Probst, K, Liu, Y, Krema, M, Fano, A. Text mining for product attribute extraction. *ACM SIGKDD Explorations Newsletter* 2006; *8*(1): 41-48.
[11] Yu, J, Zha, ZJ, Wang, M, Chua, TS. Aspect ranking: identifying important product aspects from online consumer reviews. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* 2011; 1:1496-1505.
[12] Mukherjee A, Liu B. Aspect extraction through semi-supervised modeling. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics* 2012:339-348.
[13] Chen Z, Mukherjee A, Liu B. Aspect extraction with automated prior knowledge learning. *Proceedings of ACL* 2014; 347-358.
[14] Putthividhya, D, Hu, J. Bootstrapped named entity recognition for product attribute extraction. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* 2011; 1557-1567.
[15] Chen D, Manning, CD. A fast and accurate dependency parser using neural networks. *Proceedings of EMNLP* 2014;
[16] Ratinov L, Roth, D. Design challenges and misconceptions in named entity recognition. *Proceedings of the Thirteenth Conference on Computational Natural Language Learning* 2009; 147–155.
[17] Lv, X, El-Gohary, NM. Semantic annotation for supporting context-aware information retrieval in the transportation project environmental review domain. *J. Comput. Civ. Eng*. 2015; in press.